Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   The Summer and Fall seasons see a higher demand of bikes. This is corroborated by the higher demand seen during the months from May until October.

   This is also supported by the higher demand when the weather is favorable i.e. mostly clear.

   This indicates that the demand for the bikes is higher when the weather is favorable (as per US weather patterns).

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   The default value of drop_first is False. Here a dummy variable will be created for all the levels of the categorical variable. Since we need n-1 dummy variables to be created, we specify drof_first=True. In other words, if a dummy variable is created for n-1 levels, we will be anyway knowing what the n^{th} level would be and hence we don't need a dummy variable for the same.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Temp has the highest correlation with cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   This was done by evaluating the model by plotting the scatter plot between y_test and y_pred values i.e. checking if the predicted values based on the test values exhibit a linear relationship.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   The top 3 features are:
   Temperature (coeff: 0.533595)
   Weather (Light; coeff: -0.254779)
   Year (coeff: 0.231452)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   The steps followed for linear regression are:

   1. *Getting data ready and understanding data*
      a. *Reading, understanding/visualizing*
      b. *Checking and cleaning up data (missing values, nulls, etc.)*
      c. *Evaluating and preparing data (dropping redundant cols, using appropriate categorical variables and dummy variables; with modification where needed);*
      d. *Applying EDA principles (finding relation between data using univariate, bivariate analysis)*
      e.
   2. *Building the model*
      a. *Deciding to use RFE and/or manual model building*
      b. *splitting the data into training and test data sets*
      c. *rescaling the features if/as needed*
      d. *feature selection using R-Sqaure and VIF values (RFE and Statsmodel combination to be used)*
   3. *Perform residual analysis and predictions using test data*
   4. *Evaluating the model*

2. Explain the Anscombe's quartet in detail. (3 marks)

   Anscombe's quartet is a set of four data sets, with nearly same statistical parameters, that helps us visualize data before applying any algorithms to build models.

3. What is Pearson's R? (3 marks)

   Pearson's R or Pearson correlation coefficient (PCC) helps measures linear correlation between the dependent and independent variables. A positive value indicates positive correlation and a negative value indicates negative correlation. It varies between -1 and 1. 0 indicates there is no correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Scaling helps in normalizing data between two boundary values. This is needed to bring the data, that might be having high magnitude of variation, into manageable levels.
   Normalization or minmax scaling helps bring the values between 0 and 1
   Standardization modifies the data such that it is normally distributed with mean as zero and standard deviation as 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

This happens when there is high multicollinearity between the variables. This means some of the variables may not be independently influencing the model i.e. they would be non-significant.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot helps identify if two samples of data were taken from the same source population dataset. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It helps decide if the two samples have the same distribution shape.